ANALYTICS IN ACTION

MS&E 235 (3 units)

SYLLABUS – Fall 2015

Professor	Irad Ben-Gal
Office; Hours	Tuesday 1:30pm-2:30pm
Email / Office	ibengal@stanford.edu Office: Huang MS&E #352
Classroom	Y2E2 111
Class time	Tue, Thu 12:00 PM - 1:20 PM
First : Last Class	01/04/2016 - 03/11/2016
Course Assistants CA Office Hours	rahul makhijani rahulmakhijani19@gmail.com Office hours: Mondays from 6:00 pm – 8:00 pm in Huang (Details to be posted on Canvas and Piazza), Irad

1. The Challenge

Today is the era of Big Data and Analytics. More than 300 million pictures are uploaded every day to Facebook these days and more than 100 thousands hours of video are uploaded to Youtube during the same timeframe. This is huge amount of data. If one adds to this the massive amounts of data created everyday by modern enterprises, governmental agencies, the WWW, the smartphones, the social networks and others, we end up with unprecedented amount of data. And this is only the beginning. The Wilkon Group estimates that the market value of big data will grow from $5.1B in 2012 to more than $50B in 2017[1]. A new "breed" of workers, data scientists, is now required to access those mountains of data and retrieve the relevant knowledge for decision making. The Harvard Business Review has declared data science as the "sexiest job of the 21st century[2]. The McKinsey group estimates that by 2018 there will be a shortage of 140,000 - 190,000 data scientists in the US alone as well as 1.5 million managers with the know-how to apply Big Data analysis for making effective decisions[3]. Mastering analytics, data mining and business intelligence (BI) will inevitably be key factors in harnessing Big Data for business decisions. The new Operational Intelligence Platforms which are emerging recently will create additional needs for analytics services[4]. Gartner estimates that the worldwide business intelligence (BI) software revenue will reach $13.8 billion in 2013, a 7 percent increase from 2012, reaching $17.1 billion by 2016[5].

The main focus of data science so far has been on BI which is basically a reactive process concerned with processing and organizing past data, cleaning it, editing it and presenting it in tabular and visual forms. But to affect decisions, we need a proactive process that uses the past data to *predict* future outcomes that one can base the decisions upon, for example the probability that a customer responds to a given solicitation, the likelihood of churn, the risk level of a loan applicant, and many more such applications. This is where business analytics comes in.

---

[1] http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues

[2] Data Scientist: The sexiest Job of the 21st Century, Harvard Business Review, October 2012

[3] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[4] http://www.gartner.com/DisplayDocument?ref=clientFriendlyUrl&id=2418415

[5] http://www.gartner.com/newsroom/id/2340216

Business Analytics has evolved in recent years to provide answers to these challenges. Business analytics is a new generation of computerized technologies for discovering knowledge hidden in big data. It is an interdisciplinary field lying at the interface of Statistics, Computer Science and Artificial Intelligence. Industries affected by data mining and business analytics include marketing, CRM, finance and insurance, retail, telecommunication, health care, hospitality, gaming and in fact any data-rich industry. Business applications include marketing and sales, understanding customer behavior, manufacturing processes, fraud detection, loan approval, portfolio trading, deviation detection, pattern recognition, behavioral targeting, supply chain, and more. Understanding the role of business analytics in decisions making and the underlying technologies for analyzing data will be the focus of our course.

2. **Course Overview**

The goal of the course on Business Analytics is to bridge the gap between theory and practice and empower the students to apply analytic solutions in real world applications. The course will be practically-oriented, emphasizing the knowledge discovery process, economic considerations and implementation issues.
We will study the fundamental principles and techniques of analytics, data mining and machine learning. We will examine how data analysis technologies can be used to improve decision-making. We will use real-world examples and cases to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. Homework assignments, involving realistic data (whenever possible), will provide the students with hands-on experience in applying business analytics in practice.

After taking this course you should:

1. *Approach business problems data-analytically.* Think carefully & systematically about whether & how data can improve business performance, to make better-informed decisions for management, marketing, investment, etc.

2. *Be able to interact on the topic of business analytics.* Know the fundamental principles of data science that are the basis for data mining processes, algorithms, & systems. Understand these well enough to work on data science projects and interact with everyone involved. Envision new opportunities.

3. *Have had hands-on experience mining data by using R software.* Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies.

3. Focus and interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and some technical details of data science and business analytics. The emphasis will be on understanding the fundamental concepts of data science and the business applications of data mining. This is not an algorithms course. Yet, we will provide some technical background and intuition behind some of the algorithms.

You are expected to be prepared for class discussions and understand what we have done in the prior classes. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures/discussions on the fundamental material, discussions of business applications of the ideas and techniques, case discussions, student exercises and presentations, guest lectures and demos.

Make sure you attend the class sessions, arrive prior to the starting time, remain for the entire class, and follow basic classroom etiquette, including (unless otherwise directed) having all electronic devices (that are not used for the class) turned off and put away for the duration of the class and refraining from chatting or doing other work or reading during class. In general, we will follow MS&E policies unless stated otherwise.

The Class website in for this course will contain lecture notes, reading materials, assignments, and optionally also late-breaking news. The lecture notes will be posted before the class, but the final version might be modified according to the course advancement and will be posted after the class. You should check the site daily and read all announcements and class discussion. The website address is: https://web.stanford.edu/group/canvas/discovery/

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, come to office hours to see me or the CAs, or, even better, pose your question on the discussion board (piazza) so that everyone may benefit from the answers. Also, please try to answer your classmates' questions. In grading your class participation I will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

Worth repetition: It is your responsibility to check the class website (and your email) at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent. I will make sure to check my email at least once a day.

4. Readings and Lecture Notes

Textbook: The textbook for the class will be:

- *Data Science for Business: Fundamental principles of data mining and data analytic thinking Provost & Fawcett (O'Reilly, 2013).*

http://data-science-for-biz.com/

This book covers the fundamental material that will provide the basis for you to think and communicate about data science and business analytics. At times, I may deviate from the book to cover topics in more depth or give further examples. I will complement the book with discussions of applications, cases, and demonstrations.
Other recommended books and articles are:

Optional Books:

- Eric Siegel (Author), Thomas H. Davenport Predictive Analytics: The Power to Predict: Who Will Click, Buy, Lie, or Die, 2013
- Linoff Gordon and Berry Michael, "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management", 3rd Edition, John Wiley & Sons, 2011
- Thomas Davenport & Jeanne Harris "Competing on Analytics: The New Science of Winning" 2007, Harvard Business School Press

Optional Articles:

- Xindong Wu et al., (2008), Top 10 Algorithms in Data Mining Knowledge and Information Systems, 14, 1: 1-37.
- Xindong Wu et al., (2008), 18 Candidates for the Top 10 Algorithms in Data Mining by Knowledge and Information Systems.14(1): 1-37.
- Hsinchun Chen et al., (2012). Business Intelligence and Analytics: From Big Data to Big Impact by MIS Quarterly – Special Issue: Business Intelligence Research. 36(4): 1165-1188.
- Pearl, J. & Russell, S. (2000), Bayesian networks, available at http://ftp.cs.ucla.edu/pub/stat_ser/R277.pdf
- Pedro Domingos (2012). A Few Useful Things to Know about Machine Learning CACM. 55(10). http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Lecture notes: Lecture notes for most classes will be posted on the course site few hours ahead. For some classes, I may add an appendix to include more advanced material that, if time permits, will be covered in class. Please ask questions about any material in the notes that is unclear after our class discussion and reading the book. Having the book frees up class time for more discussion of applications, cases, etc.—so many of the questions may be addressed in the book. Depending on the direction our class discussion takes, we may not cover all material in the class notes for any particular session. If the notes and the book are not adequate to explain a topic we skip, please ask about it on the discussion board or approach me at office hours.

I may hand out or post some additional required readings as we go along. *Note that some of these readings may be accessible for free only from a Stanford computer. If you can't access a link from home, please try it from school.*

For those interested in going further, please see above list of supplemental books, articles and other material (indicated in the lecture notes), which give alternative perspectives on and additional details about the topics we cover. All these are completely optional; you will not be required to know anything in these readings that are not in the primary materials or lectures.

Finally, I plan to invite a couple distinguished guest lectures from the industry to describe real-world applications of analytics. A few case studies and hands on laboratories will also be introduced throughout the semester. Be sure to attend these lectures and provide the full respect these guests deserve.

5. Requirements and Grading

The grade breakdown is as follows:

1. Homework: 50%
2. Term Project: 50%
3. Optional Participation/Professionalism/Attendance/Contribution – a correction up or down of up to 10% of the final grenade based on homework and term project

At Stanford MS&E we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum.

Homework Assignments
The homework assignments are listed (by due date) in the class schedule below. Each homework comprises questions to be answered and/or hands-on tasks. Except as explicitly noted otherwise (see next paragraph), you are expected to complete your assignments on your own—without interacting with on the completion of your assignment. You are free of course to discuss the concepts with your classmates, and to discuss similar problems to the ones in the homework.

For the hands-on parts of the assignments, you are encouraged to work with your group members and other classmates to understand how to use R to achieve what you need to do. But then you are expected to complete the assignment on your own. With the help of the CA, your classmates and myself you will surely get the support needed to cope with the course materials and homework.

The homework assignments will posted on Canvas by the CA. They are listed, by due date, in the class schedule.

Completed assignments must be typed and handed on Canvas by midnight of the submission date (that is, by 11:59pm the day before class), unless otherwise indicated. Late assignments will have their grades reduced. Answers to homework questions

should be well thought out and communicated precisely and professionally, avoiding sloppy language, poor diagrams, and irrelevant discussion.

Late Assignments: Assignments late by 24 hours will have the grade reduced by 25 %. Assignments late by 48 hours will have grade reduced by 50% and later than 48 hours will not be accepted.

Generally the Course Assistant should be the first point of contact for questions about any issue with the homework. The course assistant will have the responsibility to make sure that all questions are answered in a timely fashion. If the CA cannot help you to your satisfaction, please do not hesitate to come see me.

The hands-on tasks in the homework will be based on data that we will provide, or you will have to find. In some cases you will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class. You will use these data to build and evaluate predictive models.

For the hands-on assignments you will need to use data mining computer software. In this course we will provide support for the award-winning and free toolkit R. website for the self-tutorial for R: http://tryr.codeschool.com/

Optionally, you could try Python and its data science/analytics/visualization libraries, or use WEKA (http://www.cs.waikato.ac.nz/ml/weka/). Note however that these package won't be supported by myself or the TA.

IMPORTANT: ***You must have access to a computer on which you can install software.*** *You should bring your computer to class specifically to "Hands-on" Labs*. In these "lab session" the CA will aid anyone who needs help with configuring the software, getting it running, and dealing with the inevitable glitches that a few of you might experience. If you need additional help with using the software, please see the Course Assistant(s).

Term Project

Term project report and presentation will be prepared by teams of three students. Each team should select a business problem as well as find the relevant dataset to be analyzed. You are required to submit a proposal on the problem selected (please start looking for available data now!). Teams are encouraged to interact with the CA electronically or face-to-face in developing their project reports. You will submit various milestone deliverables through the course (see schedule). Work on the term project will be conducted throughout the semester, and specifically during the last 4 weeks of the course. The final presentations will be delivered by the students in the last 3 sessions of the course and graded by myself and the CA. These grades will be influenced by classmate evaluation mechanism that will be published later. We will further discuss the project requirements in class.

Participation/Professionalism/Contribution/Attendance

Please see Section 3.

Re-grading

If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to the CA describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., pages in the book, a copy of class notes, etc.). If the CA answer did not satisfy your claims, please send me an email with all the required material. I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go "in your favor" and possibly some go against. In fairness to all students, the entire assignment will be reconsidered.

For Students with Disabilities: If you have a qualified disability and will require academic accommodation during this course, please follow the MS&E instructions and provide me with a letter from the secretary verifying your registration and outlining the accommodations they recommend.

Class Schedule

Might change the during course

Class Number	Date	Topics (subject to change as class progresses)	Readings	Deliverables
1	Jan. 4	Introduction to the Course: "Analytics in Action"	Ch. 1	
2	Jan. 7	Infrastructure and Technologies, distributed systems	Ch. 2	Info Sheet (online)
3	Jan. 12	Data Mining Process, Supervised Learning, Classification DT Bring your laptops Short "Hands-on" Lab #1: R Basics & Data preparation	Ch. 3	
4	Jan. 14	**Guest: Big Data Infrastructure** **Dr. Navin Budhiraja** SVP, Head - Architecture & Technology at Infosys	paper	HW#1 due (R: Data preparation, visualization, etc.)
5	Jan. 19	**"Hands-on" Lab #2** Bring your laptop **Modeling (Classification/ DT/ SVM) with R**		Project Teams Formation
6	Jan. 21	Supervised Learning (cont.), SVM, Model Evaluation, **Guest: Analytics in Automotive Industry** **Dr. Ross Morrow** Senior Analytics Scientist at Ford Motor Company	Ch. 4 (p. 81-94)	HW#2 due (Classification, DT)
7	Jan. 26	Predictive Modeling Discriminant analysis, regression models	Ch. 4 (p. 94-109) Ch.5	

Class Number	Date	Topics	Readings	Deliverables
8	Jan. 28	Model performance & Analysis **Bring your laptop** **"Hands-on" Lab #3**	Ch. 5 Ch. 8	HW#3 due (SVM/S. models)
9	Feb. 2	Unsupervised Learning: Similarity Neighbors and Clusters	Ch. 6 (part 1)	**Project Proposal**
10	Feb. 4	Clustering, Collaborative Filtering, Association Rules, **Guest: Mr. Yigal Elbaz** **VP Ecosystem & Innovation at AT&T**	Ch. 6 (part 2)	HW#4 due (Regression, Regularized Models)
11	Feb. 9	Visualization and Model Performance	Ch. 7	
12	Feb. 11	Text Mining, **"Hands-on" Lab #4: unsupervised Learning**	Ch. 10	HW#5 due (Clustering, KNN, K means)
13	Feb. 16	Bayesian Belief Networks/ Causal modeling	Ch. 9	
14	Feb. 18	Link Analysis / Social nets influencers Bring your laptop **"Hands-on" Lab #5: Social nets**	Ch. 11	HW#6 due (Text Mining, PCA)
15	Feb. 23	**Guest 4: Dr. Michael Manzano** **VP Analytics Sciences (Symantec, Starbucks, GAP)**		
16	Feb. 25	Other Data Science Tasks and Techniques	Ch.12	HW#7 due (Networks)
17	March 1	Toward Analytical Engineering **Guest 5 Dr. Satya Ramaswamy, Vice President & Global Head of TCS Digital Enterprise**	Ch.13 Ch. 14	
18	March 3	Projects presentations + discussion		
19	March 8	Projects presentations + discussion		
20	March 10	Projects presentations + wrap up		Project Reports